A QUANTITATIVE EVALUATION APPROACH TO SONIFICATIONS

Katharina Vogt

Institute for Electronic Music and Acoustics - IEM
Inffeldg.10/3, 8010 Graz
Austria

ABSTRACT

Different evaluation approaches have been taken within the field of sonification. This paper introduces methods of the multi-criteria decision aid (MCDA) to the field. Different stakeholders are taken into account. In the area of explorative data analysis, the domain scientists of the field are included in the process in addition to sonification experts. The method allows to compare different sonification designs of the same or different data sets quantitatively. This enables the sonification designer to evaluate the sonification design objectively and draw conclusions on which kind of sonification is appropriate for the end user.

1. INTRODUCTION

Evaluation of sonifications is a challenging problem. On the one hand, the goal of sonification in research is exploration and gaining new insights. When this goal is reached, evaluation is no longer needed. In order to reach it, sonification methods have to be further developed, and evaluation is one important step to do this. On the other hand, 'classical' survey research can usually not be applied. It is hard to find a cohort that is large enough for statistical analysis, that is willing to really engage with first sonification approaches, and has sufficient interdisciplinary knowledge to assess the sonification for both its sonic and domain science value.

Figure 1: A group of the workshop test takers during the silent negotiation process of weighting the criteria for the MCDA.

1.1. Evaluation in the ICAD community

In a screening of all ICAD papers between 1992 and 2009 that had *evaluation* in their title, keywords and/or abstracts, I found the following meanings and concepts of evaluation within the ICAD community (as there are some 70 papers in the list, only those are cited which have evaluation as part of their title):

User Interfaces and displays: By far most of the evaluation examples stem from tests with user interfaces and displays. This category is very diverse, including the use of technical applications (from phones to cockpits [1] and train cabs [2]), sonified graphical user interfaces [3], auditory displays for visually impaired or blind people [4, 5] (e.g., auditory graphs and spreadsheets [6]), sonic interaction design [7], auditory tool palettes [8, 9], auditory icons and earcons [10], and alerts.

For these applications, efficiency assessments have been used, asking how long it takes to receive a certain information from the user interface; sometimes in comparison to other modalities, as visual display. In general, some sort of quantitative analysis is used as evaluation tool in this context.

Psycho-acoustical aspects: Psycho-acoustical aspects are evaluated. These provide insights into the relationship between stimuli and percepts, e.g. timbre or synthetized sound [11]. E.g., cross-modal influences between vision and audio [12, 13] or even multi-modal systems [14] are studied. Cross-cultural studies, cognitive factors and learning are evaluated as well. As for psycho-acoustics in general, classical auditory tests are possible [15]. Questions are, e.g., which two stimuli are different out of a set of three or how stimuli should be sorted or rated amongst each other (ABX test). (Methods of perceptual audio evaluation are discussed in [16].)

All these evaluations are done in the context of simple sounds in testing conditions, thus the methods can hardly be used for the evaluation of sonifications with complex sound phenomena involved.

Audio techniques: This category unites both audio hardware and software technology, e.g., spatial audio quality [17, 18], binaural rendering, auralisation [19] or HRTF design [20, 21]. It includes also higher levels of techniques, as semantic categorization of audio and audio-augmented reality [22].

In most of these cases, evaluations compare the objective, technical level to the psycho-acoustical level (see previous item).

Specific sonifications: A few examples of specific sonifications were evaluated, e.g., of stock market data [23]. These evaluations are often explicitly called *subjective*, e.g. in [24], because usually the sonification designer and her or his colleagues evaluate the auditory display (AD).

Others: Additionally, some papers were located in music, aesthetics or design theory and dealt mainly with the theoretical aspects of evaluation.

As for exploratory sonification exercises, conventional quantitative evaluation approaches are difficult to use. In an exploratory data display, the task is unknown and cannot be measured. Assessing the participants' individual performance does not make sense, as an exploratory interface can prove to be *good* even if it works only for one single person who obtains scientifically innovative results with it. Bovermann [25] therefore suggests a qualitative evaluation approach, based on *grounded theory* [26] and (in his case) video analysis of people using the interface. Grounded theory allows for generating hypotheses during the analysis process, in contrary to defining them beforehand. Several examples of tangible auditory interfaces were evaluated with this approach and findings are discussed by Bovermann.

1.2. Quantitative vs. qualitative evaluation

In the evaluation approaches mentioned above, an objective comparison of the different sonifications cannot be achieved. Nevertheless, this would be preferable, especially if sonifications of completely different data sets are taken into account. More general rules for successful sonification design might be deduced from the comparison of such diverse examples. The sonification designer is usually the only, but at least the primary tester. S/he surely has most expertise, but the final users of the sonification are others – in exploratory data analysis, the domain scientists. For various reasons, sonifications are often not used in scientific routines. In order to better understand the problems in the specific context, the grounded theory approach cited above can give qualitative insights. Quantitative methods are also needed, as the outcome is more measureable and for practical reasons: the effort is comparatively smaller, larger test groups can be taken into account than in the qualitative approach, and the results focus on the aspects in question. As a side benefit, the preoccupation with a sonification of a large group of domain scientists during an evaluation can increase the acceptance of this method in general.

1.3. Workshop design

This paper suggests a new method for evaluating sonifications which was tested and adapted to sonification in the workshop Science by Ear II (SBE2) that took place at the Institute for Electronic Music and Acoustics in Graz in February 2010. A group of domain scientists and sonification experts convened to elaborate sonification designs for four data sets from different scientific disciplines. In each of the four sessions, two or three teams worked in parallel on the same data set. Then, the sonification design of each team was presented in front of all workshop participants, and evaluated according to the procedure described in this paper.

2. MULTI-CRITERIA DECISION AID

During the workshop we tried a new evaluation method for ADs. The Multi-Criteria Decision Aid (MCDA, [27]) has been developed for political and economic contexts where different options need to be assessed and several criteria play a role. It may be the case that some criteria have trade-offs between each other (e.g., the need for energy supply in our society vs. an increasing ecological awareness). The MCDA incorporates consensus methods that communicate between different groups (in the economic context, these are stakeholders). For an overview of MCDA methods in the context of sustainable development see [27].

We used one method of MCDA, the *weighted sum* approach. For the context of the workshop, each sonification approach was one option O, to be rated. There were 11 sonifications in total, made of 4 categorically different data sets, each sonified by 2 or 3 different groups. The stakeholders were the participants of the workshop, i.e. domain scientists from physics or related subjects, sonification experts, and media-artists. A set of criteria c_i was established and will be discussed in detail below. These criteria were ranked according to their importance, individually and in a group process, and weights w_i were calculated for each criterion (see below). This set of weighted criteria was kept constant for the workshop, but the analysis of the results showed that it partly led to misunderstandings. Therefore I suggest a slightly refined set of criteria at the end of this section. For each sonification, each participant filled out a questionnaire and rated the AD according to each criterion. The rating r_i was averaged and multiplied by each weighting factor, then all weighted ratings were summed up to one final number W for each option:

$$W_O = \sum_i w_i \bar{r}(c_i) \tag{1}$$

During the three-day workshop, 11 sonification approaches to 4 different data sets were developed and rated. A total of 189 questionnaires were analyzed.

2.1. Set of criteria

We suggested a set of criteria at the beginning of the workshop, and this set was then extended in a discussion. Taking into account the results of the evaluation (e.g., a correlation analysis between the criteria, see Fig. 2) and feedback of the evaluation of a follow-up test (not discussed in this paper), a final set of criteria is proposed in Sec. 5. The discussed criteria were *aesthetics/amenity, intuitiveness, learning effort, clarity, potential, efficiency, 'contextability', complexity,* and *technical effort*.

The term *aesthetics* referred to the sound quality. This term was replaced with *amenity* in the discussion, as aesthetics is a broader concept from artistic research. The criterion itself was rather clear and was accepted by the participants. Amenity of the sound is important, as listeners are very sensitive to what they hear, and the level of annoyance is usually reached more rapidly than with visual displays.

Intuitiveness was one of the disputed criteria. One criticism was that intuitiveness is always achieved by learning – any AD becomes 'intuitive' after a while. The notion *familiarity* might be more appropriate in characterizing how well the sound fits the data or whether the mapping choices make sense to the listener. But the main criticism was that intuitiveness/ familiarity is not applicable to all cases, as most abstract data have no 'intuitive' sound equivalence.

Learning effort was taken into account because sonifications need to be comprehended within a rather short amount of time,

otherwise domain scientists will not start using them. The criterion was rather clear in the discussion.

Clarity refers to sounds, and how easy it is to perceive the structures of interest against a given sonic background.

Potential. Originally called benefit, this criterion was renamed potential during the workshop. A sonification shows potential if it achieves some added value, e.g., in comparison to classical displays, mathematical or numerical treatment, or for special applications. The criterion was unanimously accepted.

Efficiency was added to the criteria in the discussion, but perhaps not clearly enough defined, as discussed below. It was introduced by one of the domain scientists as a measure of efficiency in competition to classical strategies, such as visualization or mathematical treatment.

'Contextability' is a neologism for the ability of the sonification to work in certain context, defined usually in reference to the physical surrounding. Depending on the application tasks, this could be, e.g., the ability to complement a visual display, or, in laboratory condition with other sounding measurement devices, the distinctness of the AD.

Complexity was a measure of the 'non-triviality' of the sonification task. We suggested it in the first place because it is one thing if simple data are sonified (like the trend of temperature values over time), but another if 4-dimensional, highly abstract data are sonified. While the first example might be rated as a perfect sonification according to many criteria (amenity, learning effort or clarity), this is much harder for the latter. The weighted sum also needs a measure of difficulty for the task and data in order to balance the result. In the discussion this criterion was rejected for the workshop, as it measures an independent quantity - the data - , and not the sonification.

The *technical effort* was suggested by participants of the workshop. It is a measure of the applicability of a sonification, and of course it influences the probability that it is used.

	Amenity	Intuitivity	Clarity	Learning effort	Potential	Efficiency	Context-ability	Technical effort
Amenity		0,42	0,36	0,31	0,36	0,41	0,15	-0,04
Intuitivity	0,42		0,49	0,34	0,54	0,45	0,33	0,14
Clarity	0,36	0,49		0,58	0,53	0,58	0,38	-0,04
Learning effort	0,31	0,34	0,58		0,44	0,47	0,26	0,13
Potential	0,36	0,54	0,53	0,44		0,65	0,48	0,09
Efficiency	0,41	0,45	0,58	0,47	0,65		0,52	0,10
Contextability	0,15	0,33	0,38	0,26	0,48	0,52		0,20
Technical effort	-0,04	0,14	-0,04	0,13	0,09	0,10	0,20	

LEGEND:	0,4-<0,45	0,45-<0,5	0,5-<0,55	0,55-<0,6	>0,6

Figure 2: Correlation matrix showing correlation probability of pairs of criteria, calculated from all questionnaires of the workshop according to Eq. (2).

For the analysis of the criteria, a correlation matrix of all criteria pairs was calculated according to Eq. (2), where x and y are the mean samples of two matrices X and Y. The results are shown in Fig. 2. Efficiency shows dependencies with most other factors. As expected, learning effort and clarity are linked, as clearer *gestalts* are learned more readily.

$$K(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (2)$$

We also analyzed non-ratings, i.e. responses made by ticking *"don't know"* and/or *"not relevant"*. Results are shown in Fig. 3. Contextability, efficiency, and technical effort were often not rated. While efficiency was unclear (as seen from the correlation analysis), the other two criteria seemed to be only secondary and not always applicable. Potential is the only criterion that was *always* relevant, but was quite often hard to assess (*"don't know"*).

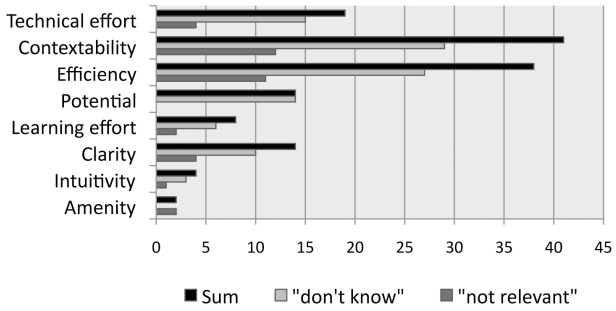

Figure 3: Number of tickings *"don't know"* or *"not relevant"* and their sum over all questionnaires of the workshop.

For the final set of criteria proposed in this paper, these analyses were taken into account.

2.2. Weighting of criteria

Normally, there are more and less important criteria. A central step of the MCDA thus is the weighting of the criteria. In the workshop, two different methods were tried out. One was the classical assessing of individual opinions by questionnaire. Every test taker had to specify percentage rates for the criteria according to their importance for the evaluation of sonifications in general. The answers were collected and averaged. The second approach was the *silent negotiation* technique of the MCDA, that achieves a consensus weighting for the whole group. All group members gather around a table, on which cards with the criteria have been placed. One side of the table is designated as "high-ranking", the opposite as "low-ranking". One by one, each person places a criterion card where s/he thinks it belongs. This procedure is repeated until no additional significant changes are made (or until a repeated pattern of changes occurs). The procedure is supervised by a moderator who does not take part in the silent negotiation, and stops the process. During the whole procedure, no discussions are allowed, but the focus on the cards allows for an intensified non-verbal 'discussion'. A photo of the silent negotiation in the workshop is shown in Figure 1.

From the ranking R_i of the silent negotiation method, weights w_i were deduced for each criterion. In a short discussion, the group agreed on a weight difference of 1:5 from the least impor-

tant to the most important criterion. The scaling is linear in our case.

$$w_i = R_i \frac{HighestRank}{NumberOfCriteria} = R_i \frac{5}{8} \quad (3)$$

The weights were ultimately normalized to the range of 0 to 1, and used to calculate the weighted sum as given in Equation 1. The resulting weights found during the workshop are shown in Fig. 4. Potential and clarity were rated first ex aequo, followed by a gap and the other criteria, the last being intuitiveness. Interestingly, averaging over the individual assessments of criteria led to very similar results which suggests the robustness of the consensus approach. Results are shown in Tab. 1 ('Average weight' vs. 'Consens weight'). Only two criteria were rated differently: intuitiveness and learning effort. The others had results within $\pm 2.5\%$ (!) of the consensus' weights. Intuitiveness was correlated with other criteria (see correlation in Fig. 2) and was extensively discussed before and during the silent negotiation (the 'intuitiveness' card was displaced demonstratively). Its final position was largely influenced by the decision of when to stop the ranking process. Learning effort was assessed as much less important in the individual rating (9.4 vs. 16.7%).

Criterium	Average Weight	Consens Weight	*Consens Re-weight*
Amenity	11,95	12,50	12,50
Intuitivity	10,11	4,17	*4,17*
Clarity	19,12	20,83	*25,00*
Learning effort	9,47	16,67	*16,67*
Potential	17,20	20,83	*25,00*
Efficiency	10,36	8,33	
Contextability	12,55	16,67	16,67
'Technical effort'	9,24	8,33	*8,33*
	100	100	*100*

Table 1: Overview of the weighting results (given in percent) comparing the two methods (average of individual questionnaires vs. consensus of silent negotiation method) and the revised set of criteria that led to slightly different weights ('Consens Re-weight').

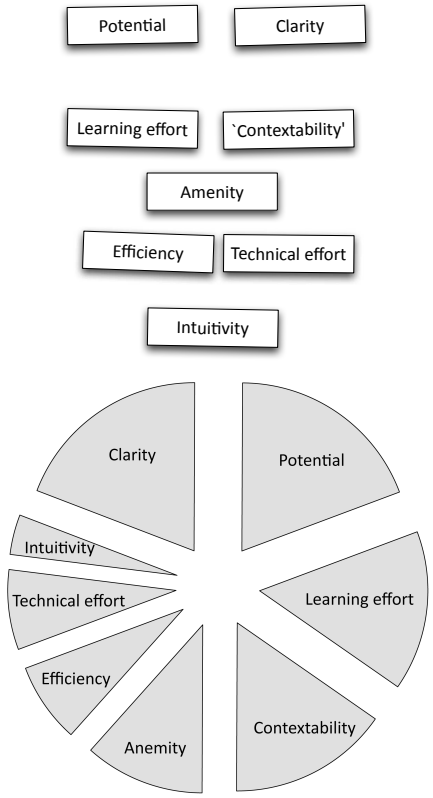

Figure 4: The ranking of the cards in the silent negotiation process in the workshop is shown in the upper part, the resulting relative weights for the criteria in the lower part of the figure. Potential and clarity were ranked highest. The normalized weights according to Eq. 3 are given as (rounded): Potential 20,8%, Clarity 20,8%, Learning effort 16,7%, Contextability 16.7%, Amenity 12,5%, Efficiency 8.3%, Technical effort 8.3%, and Intuitiveness 4.2%.

3. RESULTS FOR WORKSHOP SONIFICATIONS

The final quantitative results of the weighted sum approach are shown in Fig. 5, in chronological order. The sonifications are described in detail on the homepage http://qcd-audio.at/sbe2, but a few special examples are discussed below.

3.1. General discussion

There was a slight trend over the three days of the workshop, that ratings became generally higher. Moreover, 'threesomes' were observed for each data set, where one of the three developed sonification approaches was rated best. Thus, the three 'winning' sonifications stem from one data set each. (This must have been partly accidental, as the questionnaires were filled out immediately after the presentation of each single sonification design.)

Two more indicators are shown in Fig. 5: firstly, the difference between the rating of the sonification of the 'own team' vs. the

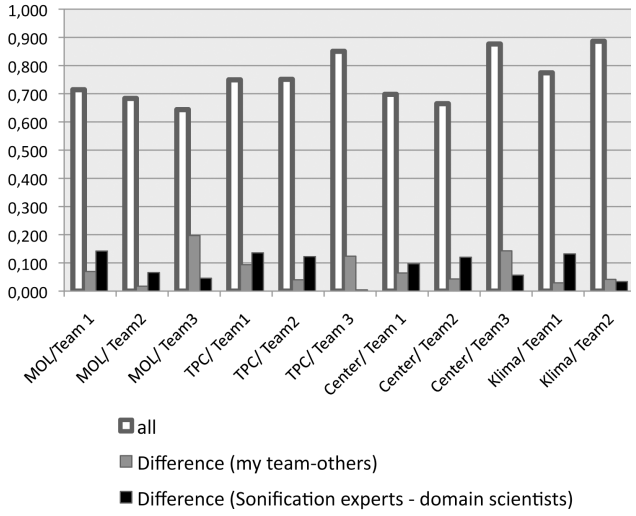

Figure 5: Weighted sums of the ratings of all sonifications developed during the workshop. The three best rated sonification approaches were 'Klima/ Team2', 'Center/ Team3', and 'TPC/ Team3', referring to the name of the data set and the number of the group. Two more comparisons are shown: firstly, the difference between how well the work of one's own team was assessed and how the members of the other groups assessed this work - the own work was always evaluated better by oneself; secondly, the difference between the ratings of the sonification experts and the domain scientists - the sonification experts generally gave better marks.

'other teams' is shown. For all sonifications, the own team was rated higher than the others. Two factors probably influenced this rating behavior. Firstly, there was often too little time to finish the sonifications properly, thus the presented results only partly reflected the real potential of the approach. Only if the idea is well understood can the real value of the sonification be assessed by people who have not been involved in its design. Secondly we noticed tendency to rate one's own work better than that of others.

The third columns in Fig. 5 show the differences between the ratings made by the sonification experts and those made by the domain scientists. In all cases, the sonification experts rated higher than the domain scientists. It can also be observed that the differences in the ratings given to the highest rated sonifications by the two groups were negligible. In general, the differences are not large and the groups rated rather homogeneously.

The overall results make it possible to compare sonifications, but a qualitative discussion of the criteria should follow. The sonifications of the workshop were developed intensively on a tight schedule. The teams had only 2-3 hours to understand the data and task, develop a sonification and implement it. Only a short period of time remained for the presentation in the plenum. Therefore in analyzing the results of the workshop, other factors besides the general criteria had to be taken into account as well. The most 'successful' sonifications were generally those whose implementation was advanced and the idea easily grasped by the other plenum members. A more thorough engagement with each sonification would be necessary for a full evaluation.

3.2. Audio examples

Authors of ICAD 2011 are encouraged, to attach sound to their papers and presentations. Therefore, we want to enable the reader with some insights into what were well and badly rated sonifications. The descriptions can neither go into detail with the data sets nor with the sonification approach.

The files can be found at http://qcd-audio.at/sbe2.

- **Climate data - The 'best' sonification vs. its direct competitor:**
  This data set consisted of climate data measured in the troposphere and stratosphere with atmospheric monthly means of temperature and refractivity, given at 9 height levels,18 latitudes, and 96 months.
  Klima_Team2_strata_allweighted.mp3:
  In the 'winning' sonification of Team 2, different height levels are mapped to different frequencies (low to high): pink noise is band pass filtered according to the data, leading to the 'windy' sound. The resulting sonification features clear rhythmic patterns for different atmospheric heights (frequency ranges). Their interdependencies can thus be studied acoustically. This sonification received high ratings throughout, including by far the most points for intuitiveness, and clearly highest for amenity. The sonification sounded like wind and thus evoked a climate metaphor, and it was also rated the least annoying sound of the whole workshop.
  Klima_Team1_pan_9levsstepwise.mp3:
  The example of Team 1 for the same data set is a sequential run through the height levels and plays the sound, as above, in time, panned in stereo for the latitudes. The result was rated less good than the one of Team 2, but still as forth best of the whole workshop.

In general, the climate data were also probably easiest to understand from all data sets of the workshop, which made the sonification simpler to design for the teams, and the ideas of the sonification simpler to grasp for the other workshop participants.

- **Center data / The second best sonification vs. one of the worst rated sonifications overall**
  This data set stemmed from computational physics, and the challenge could in principle be reduced to the question, if large coherent clusters were hidden in the three-dimensional data, and if other properties of the clusters, e.g., their shape, could be interesting.
  Center_Team3_1.32.mp3:
  In the most successful sonification for this data set, Team 3 used a systematic cluster search algorithm through the data. The sonification was creative and sounded funny: while amassing sites of a cluster following neighbor by neighbor, the sonification plays in parallel. The longer the cluster is, the more rapid the search becomes (the sound becoming quicker and higher pitched), to be then suddenly stopped and restarted slowly with a new cluster. In the audio example, soon a large cluster is found, that makes the pitch rise. Then, many more small clusters are found, leading to a randomly rhythmic succession of low pitched of sounds.
  Center_Team2.mp3:
  Team 2 chose a random position in the data, and followed a cluster through a 'timbral' space, using the 3d coordinates of the data as a 'triad' of tones mapped to frequencies between 100 and 900 Hz. The sound example plays one such path following a cluster through this space, but the used synth is rather simple and the localication cue confusing.

4. REVISED SET OF CRITERIA

The correlation analysis above showed that some notions were unclear or interpreted differently by the participants.

Amenity was a clear concept, exhibiting hardly any correlations with other criteria.

Intuitiveness was much discussed. Although we claim that intuitiveness still should be a criterion for sonification design, it showed some correlation with learning effort and clarity and thus might be disregarded as an evaluation criterion. The more intuitive/familiar a display sounds, the quicker it can be learned and the clearer the *gestalts* are perceived.

Learning effort showed some correlation with intuitiveness and clarity. We wish to retain learning effort as a criterion because the term seems to be unambiguous, and the success of a sonification is influenced by the effort it takes to learn to use it.

Clarity is also related to the mapping choices (in the case of parameter mapping), and thus should be taken into account for sonification design. In the weighting of the criteria, clarity was ranked first ex aequo with potential and is an obviously important criterion for sonification.

Potential was ranked at the highest level. In a later evaluation not discussed in this paper, participants were confused by the term potential, as it can also suggest the possibility of amelioration (also by changing the sonification!). As a more univocal term, we therefore suggest *gain* for this criterion, which was suggested during the workshop as well.

The correlation analysis also showed that the notion of *efficiency* was unclear, possibly because efficiency may mean economy of time either for the sonification itself or compared to classical displays. Because the correlation matrix showed that efficiency was quite highly correlated to potential, we excluded it from the final set of criteria.

The next three criteria are rather secondary for the evaluation of sonifications, and apply only if they are appropriate: *Contextability* might not play a large role in exploratory data analysis. *Complexity* stands for non-triviality, and refers to the 'challenge' set by the data and task. The *technical effort* is ambivalent, as technical issues may be solved differently (and will become easier in the future) while the sonification design remains the same. Still, the criterion showed the lowest correlation with other factors and is all clear cut.

The proposed final set of criteria for the evaluation of sonifications is shown below, with questions defining the terms more detailed:

Gain	How much is gained by the sonification, e.g., in comparison to other displays or classical methods?
(Gestalt) Clarity	How clearly can differences and interesting structures be perceived in the sonification?
Learning effort	How long does it take to comprehend the sonification and to be able to make use of it?
(Sound) Amenity	How aesthetically pleasing (as opposed to annoying) is the sound?
Additional criteria can be added if useful:	
'Contextability':	Is the sonification applicable in its context (e.g., scientific exploration, public outreach, work environment, etc.)
(Task&data) Complexity:	How complex (or 'non-trivial') did you think the task or underlying data were (not the sonification or sound!)?
Technical effort:	How much technical effort did the sonification require?

Table 2: Criteria for evaluating sonifications.

Re-analysis of the workshop data

Because the revised set of criteria differed from the one used during the workshop, we re-analyzed the data with different weights. We omitted efficiency, which had shown high correlation with potential, and also other criteria. Furthermore, we took the gap between the cards ranked first and second into account, which was not done in the first analysis. The second card row then is given rank '3'. These modification of the weights did not change the overall result: in all different assessments, the final weights for each criterion changed only slightly, and there were hardly any effects seen in the final relative weighted sums. The revised set of criteria and the one used during the workshop are so similar that the weighted sum results of the workshop data is still valid even with new criteria.

5. CONCLUSIONS

An exploratory sonification will always be ultimately evaluated on the basis of its exploration gain, i.e. new insights in a field that have been supported or inspired by the sonification. Nevertheless, evaluation is needed until sonifications can become that successful.

The weighted sum approach has also drawbacks. The first is inherent: completely different categories, 'apples and oranges', are summed to one final number. However, this is also a distinct benefit of the method. Second, while the theoretical scale of results is one over the highest possible rating ($1/rating_{max}$) to 1, its *effective scale* seems to be much smaller. On the one hand, *some* of the criteria will always be assessed as good, and the effective minimum will lie much higher. On the other hand, hardly ever will *all* criteria receive maximal ratings (from all participants!), thus the maximum lies below 1. For the workshop, the results lay between 0.6 and 0.9, which leaves only small differences between the options.

In general it can be concluded that the MCDA is a useful method for comparing different sonifications quantitatively. It objectifies the evaluation to a certain extent. Nevertheless, a qualitative analysis of *why* some approaches are rated better than others has to follow. Such an analysis can be used to improve future sonification designs. The evaluation process itself is fruitful for the sonification designer, because it includes the domain scientists in a discourse across different criteria.

Remark.

Parts of this paper have already been published within my thesis [28].

Acknowledgments.

I would like to thank all participants of the workshop, Ines Omann for her practical insights into the MCDA, Anna Katharina Fuchs for organizational support with the workshop and the evaluation, and Hank Fullenwider for proof-reading.

6. REFERENCES

[1] D. S. Brungart and B. D. Simpson, "Design, validation, and in-flight evaluation of an auditory attitude indicator based on pilot-selected music," in *Proceedings of the 15th International Conference on Auditory Display, Paris*, 2008.

[2] P. Zwolinski and J. Sagot, "A simulation approach to the design and evaluation of audiory interfaces in a high speed train driving cab," in *Proceedings of the 5th International Conference on Auditory Display, Glasgow*, 1998.

[3] G. Wersényi, "Evaluation of auditory representations for selected applications of a graphical user interface," in *Proceedings of the 16th International Conference on Auditory Display, Copenhagen*, 2009.

[4] R. D. Stevens, S. Brewster, P. C. Wright, and A. D. N. Edwards, "Design and evaluation of an auditory glance at algebra for blind readers," in *Proceedings of the 2nd International Conference on Auditory Display, Santa Fe*, 1994.

[5] G. Wersényi, "Evaluation of user habits for creating auditory representations of different software applications for blind persons," in *Proceedings of the 15th International Conference on Auditory Display, Paris*, 2008.

[6] T. Stockman, "The design and evaluation of auditory access to spreadsheets," in *Proceedings of the 11th International Conference on Auditory Display, Sydney*, 2004.

[7] S. Pauletto and A. Hunt, "Interacting with sonifications: An evaluation," in *Proceedings of the 13th International Conference on Auditory Display, Montreal*, 2007.

[8] S. Brewster and C. V. Clarke, "The design and evaluation of a sonically-enhanced tool palette," in *Proceedings of the 4th International Conference on Auditory Display, Palo Alto*, 1997.

[9] M. Crease and S. Brewster, "Making progress with sounds - the design and evaluation of an audio progress bar," in *Proceedings of the 5th International Conference on Auditory Display, Glasgow*, 1998.

[10] P. Lucas, "An evaluation of the communicative ability of auditory icons and earcons," in *Proceedings of the 2nd International Conference on Auditory Display, Santa Fe*, 1994.

[11] M. Mellody and G. H. Wakefield, "A tutorial example of stimulus sample discrimination in perceptual evaluation of synthesized sounds: discrimination between original and resynthesized singing," in *Proceedings of the 7th International Conference on Auditory Display*, 2001.

[12] D. Devallez, D. Rocchesso, and F. Fontana, "An experimental evaluation of the influence of auditory cues on perceived visual orders in depth," in *Proceedings of the 13th International Conference on Auditory Display, Montreal*, 2007.

[13] U. Reiter and M. Weitzel, "Influence of interaction on perceived quality in audiovisual applications: Evaluation of cross-modal influence," in *Proceedings of the 13th International Conference on Auditory Display, Montreal*, 2007.

[14] D. McGookin and S. Brewster, "Dolphin: The design and initial evaluation of multimodal focus and context," in *Proc. of the 9th International Conference on Auditory Display, Kyoto*, 2002.

[15] A. Martins and R. M. Rangayyan, "Experimental evaluation of Auditory Display and sonification of textured images," in *Proceedings of the 4th International Conference on Auditory Display, Palo Alto*, 1997.

[16] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley and Sons, 2006.

[17] C. Guastavino, V. Larcher, G. Catusseau, and P. Boussard, "Spatial audio quality evaluation: Comparing transaural, ambisonics and stereo," in *Proceedings of the 13th International Conference on Auditory Display, Montreal*, 2007.

[18] H. J. Song, K. Beilharz, and D. Cabrera, "Evaluation of spatial presentation in sonification for identifying concurrent audio streams," in *Proceedings of the 14h International Conference on Auditory Display, Montreal*, 2007.

[19] T. Lokki and H. Jaervelaeinen, "Subjective evaluation of auralization of physics-based room acoustics modeling," in *Proceedings of the 8th International Conference on Auditory Display, Helsinki*, 2001.

[20] P. Mokhtari, R. Nishimura, and H. Takemoto, "Toward HRTF personalization: an auditory-perceptual evaluation of simulated and measured HRTFs," in *Proceedings of the 15th International Conference on Auditory Display, Paris*, 2008.

[21] S. Yairi, Y. Iwaya, and Y. Suzuki, "Individualization feature of head-related transfer functions based on subjective evaluation," in *Proceedings of the 15th International Conference on Auditory Display, Paris*, 2008.

[22] D. L. Jones, K. M. Stanney, and H. Foaud, "An optimized spatial audio system for virtual training simulations: Design and evaluation," in *Proceedings of the 11th International Conference on Auditory Display, Limerick*, 2005.

[23] K. V. Nesbitt and S. Barrass, "Evaluation of a multimodal sonification and visualisation of depth of market stock data," in *Proceedings of the 9th International Conference on Auditory Display, Kyoto*, 2002.

[24] W. L. Martens, "Subjective evaluation of auditory spatial imagery associated with decorrelated subwoofer signals," in *Proceedings of the 9th International Conference on Auditory Display, Kyoto*, 2002.

[25] T. Bovermann, "Tangible auditory interfaces. Combining Auditory Display and tangible interfaces." Ph.D. dissertation, University of Bielefeld, 2009.

[26] Wikipedia: Grounded theory. [Online]. Available: http://en.wikipedia.org/wiki/Grounded_theory

[27] I. Omann, "Multi-criteria decision aid as an approach for sustainable development analysis and implementation," Ph.D. dissertation, University of Graz, 2004.

[28] K. Vogt, "Sonification of simulations in computational physics," Ph.D. dissertation, University of Music and Performing Arts Graz, 2010.